IDV assignment 2

Rayyan Hassan

March 2022

1 Introduction

The document provides a static visualization on the American Community Survey dataset. It contains a variety of different features such as race, gender, annual income, work hours, education, marital status, native country, occupation, sex etc. Using various combinations from this data, it is possible to extrapolate new information which can better assist us in understanding the different aspects of the community.

1.1 Pre-Processing

It is quite common for the data to not be clean as such it is necessary to first pre process it so that the data can be used to implement the algorithm. In this case, we do not need to use the data to implement an algorithm rather we are simply performing feature engineering and are trying to better understand the different co relations or statistical anomalies that exist inside the data. To do this, I simply replaced the columns which contain '?' by the mean of the respective columns. Other approaches also existed such as taking the median, mode or even simply deleting the entries however I found this approach to be the most suitable.

2 Visualization Question

Currently, there are many reasons behind which there is a great disparity between the annual payments of the workforce in the United States. These differences are caused by a variety of different factors such as age, race, gender and even the quality of education. Thus my research question became to analyze whether there exists a significant difference in both the capital gain and the quality of education between the different ethnic groups. This could also further translate to checking whether this difference also exists in the two different genders as well. The only true concern that I had regarding this question was how reliable the data would be considering not only the different working hours that most people work but also whether historical events such as the great recession have effected the data in a biased way.

2.1 Visualization Decisions

I made some particular decisions regarding how I would implement the design and the majority of those were influenced by the 3rd lecture of IDV. I tried to balance the data by using the rule of removing the ink to improve the data to ink ratio. I ideally wanted to keep the data easily readable to new users thus I didn't want to use too much data but at the same time, I wished to keep the visualization simple as well and as such limited the use of colors as well. I did this by incorporating the legends in the data however I kept them as stand alone to better improve the data ink ratio. I also increased the number of graphs rather than report more data on the same graph so as to improve the data density as well. This I believed was quite important as it makes the charts easily interpret-able. Lastly, I also removed the grids from the images to reduce the chart junk as well.

2.2 Visualization Tools

Initially, I just plotted simple bar graphs between these different attributes in order to see if I could extrapolate any useful information and to get a feel for the data. After understanding the domain, I moved onto utilizing the data to extract the information that I needed. With my first visualization, (1), I wished to see the impact that the level of education had on the annual income of the populace. Not surprisingly, I saw that the greatest income was generally held by people who had at least graduated their university, i.e held their bachelors degree. Therefore, the best jobs were held by people who had a bachelors, masters or doctorate degree respectively. I then moved on to see the difference between the number of men and women pursuing further education considering that their population is approximately equal. My findings showed (2) that the differences between the male and female populace pursuing further education was quite staggering as barely 10% of the total doctoral candidates were women and only 20% of the total masters graduates were comprised of women as well. Thus, i was able to quite clearly see the difference in the capital gain of both males and females of the same gender which indicated the presence of a gender gap. However, just to put the nail in the coffin so to speak, I did a general overview of the difference in the annual income and the capital gain of women as well which can be seen in (4).

I did also want to do a cursory overview regarding the different races and whether the pay gap extended there as well. I checked the quality of education as the main indicator of whether they were earning up to par or not. According to my findings (3), majority of the graduates from both high school and further studies were comprised of white students. However considering the vast differences in population, it was not an accurate measure which is why I did a percentile analysis as well. The difference (5) between the white and black population for example is approximately five times however the data indicated that the difference between the high school graduates is approximately 10 times and it increases more fold the further into their studies they go